(*Supplementary File*)

Towards Reliable Online Just-in-time Software Defect Prediction

George Cabral, *Member, IEEE,* Leandro L. Minku, *Senior Member, IEEE*

Abstract

This document presents additional information to the paper "Towards Reliable Online Just-in-time Software Defect Prediction, IEEE-TSE". The tables and figures in this document are derived from the experiments carried out in the paper, but present more detailed information and/or different perspectives.

REFERENCES

[1] N. Mittas and L. Angelis, "Ranking and clustering software cost estimation models through a multiple comparisons algorithm," *IEEE TSE*, vol. 39, no. 4, pp. 537–551, 2013.

[2] G. G. Cabral, L. L. Minku, E. Shihab, and S. Mujahid, "Class imbalance evolution and verification latency in just-in-time software defect prediction," in *ICSE*, 2019, pp. 666–676.

G. Cabral was with the Department of Computing, Federal Rural University of Pernambuco, BR, george.gcabral@ufrpe.br
L. Minku was with the Department of Computer Science, University of Birmingham, UK, L.L.Minku@cs.bham.ac.uk

TABLE I: Average predictive performance results for all the experimented methods, including the proposed method PBSA.

Dataset	Classifier	rec(0)	rec(1)	|rec(0) - rec(1)|	gmean
Fabric	OOB	50.24 (2.20)[-b]	74.45 (2.15)[b]	28.50 (2.97)[-b]	59.04 (0.63)[-b]
	UOB	42.28 (5.94)[-b]	83.70 (4.68)[b]	43.91 (8.76)[-b]	57.07 (2.42)[-b]
	OOB-SW	73.32 (0.11)[b]	46.36 (0.36)[-b]	47.73 (0.32)[-b]	50.74 (0.29)[-b]
	ORB	60.35 (1.75)	68.36 (1.44)	20.59 (2.95)	60.93 (0.87)
	PBSA	66.37 (1.02)[b]	61.42 (1.64)[-b]	14.46 (0.62)[b]	61.20 (0.80)[s]
Jgroups	OOB	59.38 (1.40)[-b]	56.67 (1.52)[-b]	28.13 (1.45)[-b]	54.71 (0.50)[-b]
	UOB	73.78 (3.10)[b]	45.12 (3.02)[-b]	36.81 (3.66)[-b]	55.09 (0.71)[-b]
	OOB-SW	81.50 (0.07)[b]	35.33 (0.26)[-b]	57.51 (0.31)[-b]	47.95 (0.25)[-b]
	ORB	62.65 (0.79)	56.73 (1.23)	17.79 (1.30)	57.76 (0.96)
	PBSA	65.64 (0.56)[b]	52.94 (0.96)[-b]	16.32 (0.80)[b]	57.68 (0.55)[-b]
Camel	OOB	57.06 (1.44)[-b]	73.99 (0.92)[b]	25.47 (1.60)[-b]	62.90 (0.70)[-b]
	UOB	55.57 (2.53)[-b]	71.28 (2.34)[s]	29.27 (2.73)[-b]	60.35 (1.05)[-b]
	OOB-SW	71.67 (0.11)[b]	40.38 (0.27)[-b]	64.61 (0.31)[-b]	40.29 (0.41)[-b]
	ORB	60.74 (0.61)	70.41 (0.77)	17.03 (0.87)	63.63 (0.50)
	PBSA	68.60 (0.47)[b]	66.99 (0.67)[-b]	11.67 (0.49)[b]	66.72 (0.44)[b]
Tomcat	OOB	59.82 (2.41)[*]	61.75 (1.79)[-b]	29.42 (2.29)[-b]	57.28 (0.88)[-b]
	UOB	68.48 (2.51)[b]	50.04 (2.57)[-b]	33.69 (2.85)[-b]	55.18 (1.05)[-b]
	OOB-SW	65.20 (0.15)[b]	52.30 (0.27)[-b]	35.93 (0.25)[-b]	54.53 (0.19)[-b]
	ORB	59.43 (1.39)	64.37 (0.68)	16.08 (1.52)	60.18 (0.80)
	PBSA	66.33 (0.71)[b]	58.04 (0.92)[-b]	14.62 (0.86)[b]	61.19 (0.67)[b]
Brackets	OOB	49.11 (0.27)[-b]	89.49 (0.29)[b]	41.90 (0.25)[-b]	63.94 (0.09)[m]
	UOB	54.59 (1.52)[-b]	83.10 (1.64)[b]	32.98 (2.17)[b]	64.24 (0.44)[b]
	OOB-SW	56.29 (0.10)[-b]	79.82 (0.28)[b]	42.80 (0.20)[-b]	61.68 (0.16)[-b]
	ORB	61.68 (1.07)	77.15 (1.05)	36.01 (1.44)	63.66 (0.48)
	PBSA	65.56 (0.89)[b]	74.25 (1.00)[-b]	24.93 (1.49)[b]	65.49 (0.48)[b]
Neutron	OOB	69.71 (0.79)[-b]	91.89 (0.62)[b]	23.97 (1.33)[-b]	79.32 (0.45)[-b]
	UOB	58.83 (1.60)[-b]	92.45 (0.34)[b]	38.41 (1.54)[-b]	70.73 (1.43)[-b]
	OOB-SW	73.83 (0.10)[-b]	83.08 (0.36)[b]	20.49 (0.32)[-b]	76.74 (0.21)[-b]
	ORB	79.89 (1.63)	81.12 (1.37)	13.98 (2.14)	79.93 (0.46)
	PBSA	74.96 (0.42)[-b]	86.44 (0.49)[b]	19.53 (0.72)[-b]	79.88 (0.19)[-b]
Spring Integration	OOB	62.48 (2.06)[-b]	53.74 (1.81)[b]	47.28 (1.82)[-b]	48.12 (0.72)[-b]
	UOB	55.65 (6.14)[-b]	59.31 (3.15)[b]	37.58 (4.87)[b]	52.19 (2.78)[*]
	OOB-SW	45.55 (0.21)[-b]	79.88 (0.34)[b]	39.52 (0.30)[-b]	56.12 (0.24)[b]
	ORB	74.33 (0.86)	44.31 (1.24)	37.30 (1.73)	52.20 (0.83)
	PBSA	75.61 (0.55)[b]	43.58 (1.01)[-b]	36.31 (0.94)[b]	52.16 (0.58)[-b]
Broadleaf	OOB	59.25 (1.23)[-b]	68.33 (1.89)[m]	33.40 (2.50)[-b]	60.07 (0.75)[-b]
	UOB	59.32 (4.45)[-b]	62.69 (4.92)[-b]	43.26 (4.37)[-b]	55.46 (1.31)[-b]
	OOB-SW	78.21 (0.10)[b]	34.73 (0.31)[-b]	71.02 (0.33)[-b]	37.00 (0.60)[-b]
	ORB	61.60 (1.48)	67.00 (1.47)	19.17 (2.32)	61.97 (0.76)
	PBSA	66.48 (0.70)[b]	62.55 (3.20)[-b]	12.41 (3.36)[b]	62.65 (3.26)[b]
Nova	OOB	68.54 (0.35)[-b]	86.27 (0.70)[b]	24.34 (0.59)[-b]	75.41 (0.23)[-b]
	UOB	65.56 (0.63)[-b]	90.84 (0.89)[b]	27.60 (0.75)[-b]	75.94 (0.15)[m]
	OOB-SW	66.41 (0.05)[-b]	85.90 (0.17)[b]	33.66 (0.17)[-b]	72.85 (0.10)[-b]
	ORB	75.44 (2.26)	79.78 (4.06)	20.28 (1.51)	75.57 (0.94)
	PBSA	75.15 (0.40)[-b]	82.93 (0.47)[b]	13.63 (0.64)[b]	77.72 (0.24)[b]
NPM	OOB	37.92 (1.84)[-b]	74.89 (1.17)[b]	49.68 (1.52)[-b]	46.17 (0.77)[-b]
	UOB	38.27 (2.67)[-b]	72.83 (1.88)[b]	48.90 (1.71)[-b]	45.87 (0.94)[-b]
	OOB-SW	55.56 (0.18)[s]	62.75 (0.54)[-b]	43.68 (0.31)[-b]	50.53 (0.35)[-b]
	ORB	55.25 (0.84)	63.95 (1.31)	31.74 (1.70)	54.26 (0.91)
	PBSA	61.83 (1.12)[b]	56.11 (2.03)[-b]	16.57 (1.18)[b]	56.87 (1.45)[b]
Ranking	OOB	3	1	3	2
	UOB	3	1	4	2
	OOB-SW	1	4	5	3
	ORB	2	2	2	1
	PBSA	1	3	1	1

Standard deviations across 30 runs are shown in brackets. Symbols [*], [s], [m] and [b] represent insignificant, small, medium and big A12 effect size against ORB. Presence/absence of the sign "-" in the effect size means that the corresponding approach was worse/better than ORB. The groups' rankings with smaller numbers indicate better ranks according to Scott-Knott-A12 test [1]. The average predictive performances and A12 effect sizes of OOB, UOB, OOB-SW and ORB are from [2].

TABLE II: Standard deviations through time for all the experimented methods, including the proposed method PBSA.

Dataset	Classifier	rec(0)	rec(1)	|rec(0) - rec(1)|	g-mean
Fabric	OOB	16.79 (0.67)[-s]	11.60 (0.68)[b]	20.33 (1.24)[b]	11.96 (0.37)[b]
	UOB	16.41 (1.25)[*]	10.29 (2.32)[b]	19.63 (2.19)[b]	12.51 (0.37)[b]
	OOB-SW	25.34 (0.07)[-b]	27.14 (0.23)[-b]	30.46 (0.15)[-b]	16.53 (0.20)[-m]
	ORB	15.31 (2.58)	18.35 (0.56)	22.75 (1.84)	16.08 (0.67)
	PBSA	11.30 (0.41)[b]	16.89 (0.38)[b]	20.52 (0.19)[b]	15.32 (0.25)[b]
Jgroups	OOB	22.24 (0.54)[-b]	15.26 (0.74)[-b]	20.22 (0.47)[-b]	11.82 (0.31)[-b]
	UOB	18.32 (1.62)[-b]	14.61 (0.85)[-b]	19.58 (1.03)[-b]	10.61 (0.49)[s]
	OOB-SW	22.58 (0.12)[-b]	22.09 (0.24)[-b]	25.58 (0.15)[-b]	12.77 (0.15)[-b]
	ORB	14.09 (0.83)	13.36 (0.58)	17.36 (0.64)	10.71 (0.51)
	PBSA	11.45 (0.35)[b]	11.28 (0.48)[b]	16.22 (0.40)[b]	9.66 (0.22)[b]
Camel	OOB	18.84 (0.68)[-b]	11.48 (0.70)[b]	21.15 (0.85)[-b]	10.21 (0.43)[b]
	UOB	19.98 (1.17)[-b]	13.58 (1.28)[-b]	20.75 (1.65)[-b]	9.55 (0.85)[b]
	OOB-SW	31.06 (0.11)[-b]	33.81 (0.21)[-b]	29.26 (0.15)[-b]	17.28 (0.39)[-b]
	ORB	14.58 (0.63)	11.87 (0.37)	19.45 (0.91)	10.80 (0.31)
	PBSA	8.86 (0.38)[b]	11.39 (0.23)[b]	13.57 (0.31)[b]	9.42 (0.13)[b]
Tomcat	OOB	20.07 (1.10)[-b]	18.02 (1.57)[-b]	21.49 (1.61)[-b]	10.72 (1.07)[-b]
	UOB	19.83 (1.82)[-b]	16.89 (1.60)[-b]	21.35 (1.88)[-b]	9.77 (1.60)[s]
	OOB-SW	23.91 (0.09)[-b]	17.99 (0.20)[-b]	21.71 (0.14)[-b]	9.28 (0.10)[b]
	ORB	15.48 (0.95)	10.00 (0.62)	18.04 (1.12)	9.66 (0.53)
	PBSA	10.98 (0.59)[b]	8.93 (0.57)[b]	12.37 (0.55)[b]	7.49 (0.28)[b]
Brackets	OOB	14.80 (0.36)[b]	9.47 (0.48)[b]	18.01 (0.27)[b]	15.37 (0.06)[b]
	UOB	18.09 (0.85)[-b]	13.69 (0.97)[b]	23.98 (0.79)[-s]	15.55 (0.13)[b]
	OOB-SW	22.77 (0.07)[-b]	22.88 (0.21)[b]	24.54 (0.13)[-b]	16.43 (0.11)[b]
	ORB	16.25 (0.31)	25.29 (0.26)	23.81 (0.86)	18.57 (0.19)
	PBSA	11.27 (0.21)[b]	22.35 (0.27)[b]	22.77 (0.48)[b]	18.98 (0.15)[-b]
Neutron	OOB	12.00 (0.96)[-b]	8.65 (0.40)[b]	13.20 (1.06)[-b]	9.66 (1.29)[-b]
	UOB	21.76 (1.31)[-b]	12.32 (0.43)[m]	21.25 (1.13)[-b]	18.64 (1.37)[-b]
	OOB-SW	15.22 (0.11)[-b]	13.78 (0.52)[-b]	17.60 (0.25)[-b]	11.57 (0.20)[-b]
	ORB	7.42 (0.27)	12.51 (1.63)	11.34 (0.55)	6.14 (0.38)
	PBSA	8.58 (0.30)[-b]	13.07 (0.55)[*]	9.91 (0.50)[b]	6.81 (0.26)[-b]
Spring Integration	OOB	30.72 (1.08)[-b]	28.71 (0.68)[-b]	30.97 (1.11)[-b]	19.21 (0.56)[-b]
	UOB	25.14 (1.67)[-b]	20.39 (2.12)[b]	21.78 (2.45)[b]	12.94 (1.06)[b]
	OOB-SW	24.10 (0.17)[-b]	14.69 (0.33)[b]	29.34 (0.24)[-m]	16.29 (0.17)[b]
	ORB	18.74 (0.43)	21.39 (0.67)	28.91 (0.93)	17.57 (0.59)
	PBSA	17.38 (0.62)[b]	23.34 (0.64)[-b]	32.12 (0.65)[-b]	18.33 (0.41)[-b]
Broadleaf	OOB	23.67 (0.97)[-b]	16.05 (1.58)[-b]	20.08 (0.74)[-b]	11.41 (0.33)[b]
	UOB	25.06 (1.56)[-b]	23.03 (2.28)[-b]	20.04 (2.04)[-m]	10.18 (0.75)[b]
	OOB-SW	28.78 (0.12)[-b]	35.66 (0.29)[-b]	29.43 (0.26)[-b]	19.65 (0.53)[-b]
	ORB	15.01 (1.03)	14.49 (0.69)	18.74 (0.80)	12.64 (0.22)
	PBSA	10.27 (0.37)[b]	12.52 (0.31)[b]	16.09 (0.19)[b]	11.64 (0.14)[b]
Nova	OOB	16.20 (0.27)[-b]	12.17 (0.72)[b]	16.47 (0.39)[b]	12.54 (0.28)[b]
	UOB	14.84 (0.35)[b]	9.33 (1.15)[b]	15.66 (0.53)[b]	12.64 (0.17)[b]
	OOB-SW	17.93 (0.06)[-b]	19.49 (0.29)[-b]	17.83 (0.11)[s]	14.58 (0.09)[-b]
	ORB	15.13 (0.90)	15.57 (0.85)	17.93 (0.68)	13.96 (0.64)
	PBSA	9.15 (0.22)[b]	15.16 (0.28)[b]	14.50 (0.17)[b]	13.46 (0.11)[b]
NPM	OOB	27.86 (1.24)[-b]	18.41 (1.06)[b]	29.56 (1.64)[-b]	16.67 (0.81)[-b]
	UOB	28.60 (2.05)[-b]	19.00 (1.19)[b]	28.69 (1.29)[-b]	16.36 (0.49)[-b]
	OOB-SW	32.84 (0.09)[-b]	21.57 (0.30)[-b]	29.90 (0.11)[-b]	18.15 (0.26)[-b]
	ORB	22.87 (0.87)	19.98 (0.74)	26.96 (1.03)	14.80 (0.53)
	PBSA	13.28 (0.52)[b]	13.20 (0.64)[b]	17.20 (0.34)[b]	11.88 (0.46)[b]
Ranking	OOB	3	1	2	2
	UOB	3	1	2	2
	OOB-SW	4	3	5	3
	ORB	2	2	2	2
	PBSA	1	1	1	1

Standard deviations across 30 runs are shown in brackets. Symbols [*], [s], [m] and [b] represent insignificant, small, medium and big A12 effect size against ORB. Presence/absence of the sign "-" in the effect size means that the corresponding approach was worse/better than ORB. The groups' rankings with smaller numbers indicate better ranks according to Scott-Knott-A12 test [1].

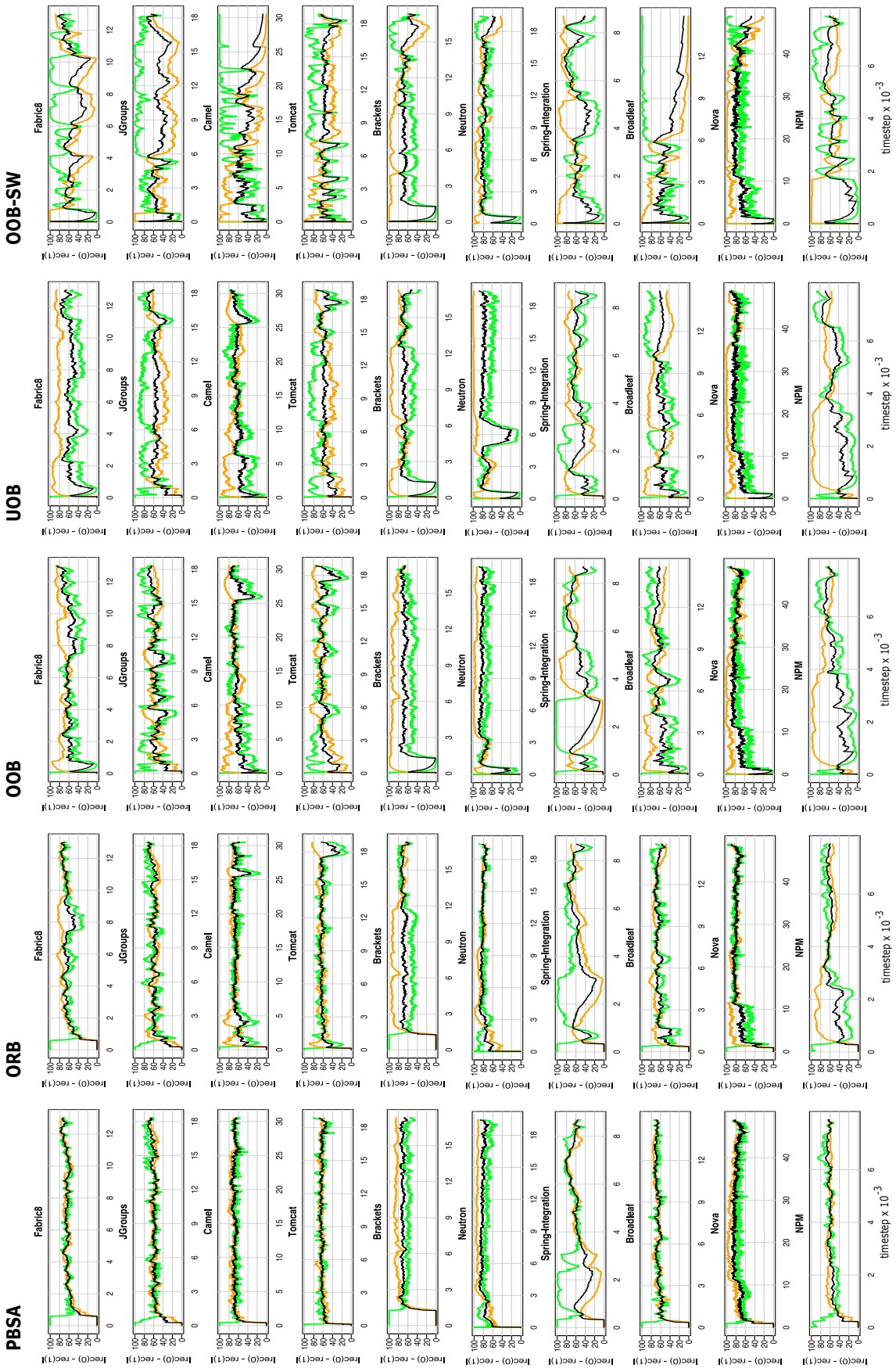

Fig. 1: Average performance through time for the all the methods. (i) the black lines represent the g-mean; (ii) the green lines represent the recall of clean class ($rec(0)$); (iii) and the orange lines represent the recall of the defect-inducing class ($rec(1)$).

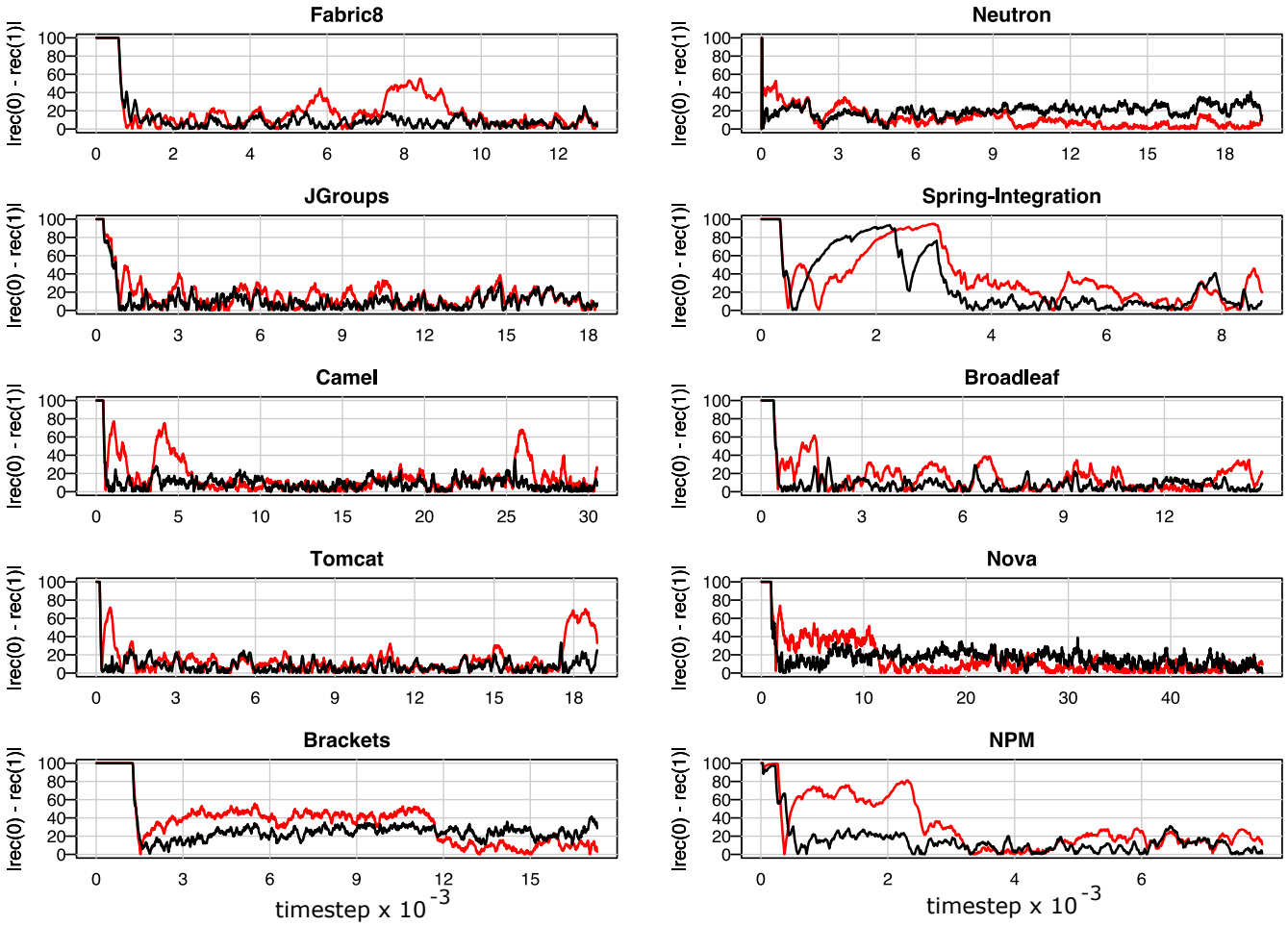

Fig. 2: Difference between the recalls $\|rec(0) - rec(1)\|$ over time for the methods ORB (red) and PBSA (black) for each dataset.

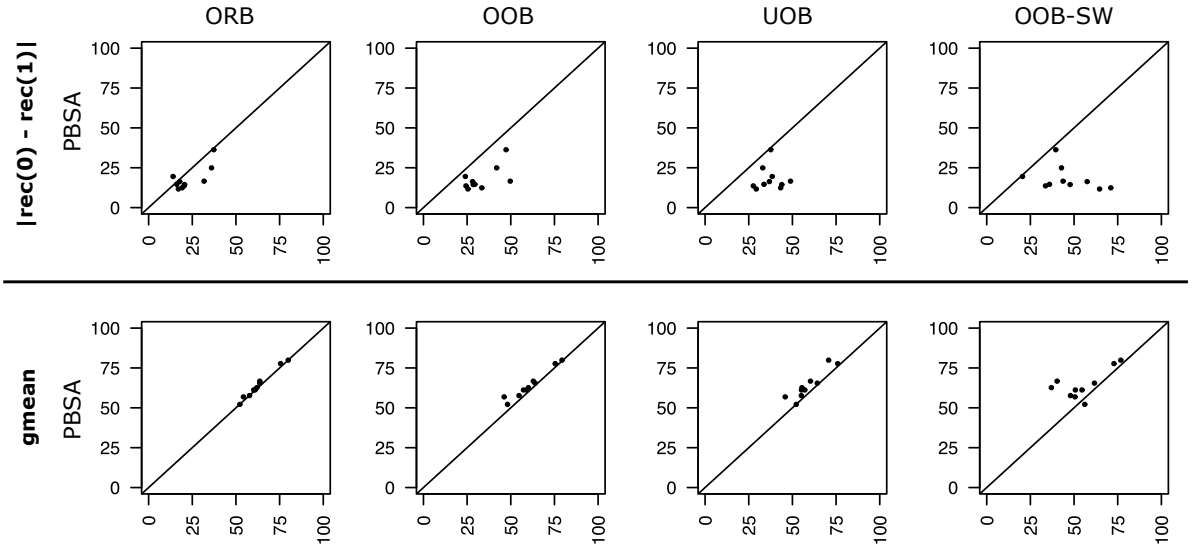

Fig. 3: Paired comparison in terms of difference between the recalls $\|rec(0) - rec(1)\|$ and g-mean between PBSA and the remaining methods. Each point corresponds to one of the ten datasets. For plots in the first row ($\|rec(0) - rec(1)\|$), points below the diagonal line indicate a better performance for PBSA. For plots in the second row (g-mean), points above the diagonal line indicate a better performance for PBSA.